Optimizing Gender Classification in Speech Data Using Hyperparameter-Tuned CNN-LSTM Models

Sarihaddu Kavitha[1], B. Basaveswara Rao[2], Suneetha Bulla [3]

[1]Research Scholar, Acharya Nagarjuna University, Assistant Professo, , Koneru Lakshmaiah Education Foundatio , kavitha.sarihaddu@gmail.com
[2]Professor, Acharya Nagarjuna University , bobbabrao62@gmail.com
[3]Associate Professor, Koneru Lakshmaiah Education Foundation, Vaddeswaram, suneethabulla@gmail.com

This paper proposes optimizing the performance of a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model for gender identification from speech by implementing Hyperparameter Tuning. The CNN-LSTM model, which effectively captures spatial and temporal features in speech data, can be further enhanced by fine-tuning key hyperparameters. We explore techniques like Grid Search and Random Search to systematically search for the best combination of hyperparameters, including the number of CNN filters, LSTM units, learning rate, batch size, and dropout rate. By optimizing these hyperparameters, the model's ability to generalize and distinguish gender-based variations in speech is improved, leading to higher accuracy, precision, recall, and F1-score. Experimental results show that hyperparameter tuning significantly boosts the model's performance compared to default settings, providing a more robust and efficient framework for gender classification. This approach offers valuable insights into improving model performance and can be extended to other speech and audio-based classification tasks.

1. Introduction

Speech is one of the most natural forms of human communication, and it carries a wealth of information beyond just the content of what is being said. Among these features, gender identification plays a critical role in various applications, including personalized voice-activated systems, telecommunication services, demographic analysis, and human-computer interaction. Accurate gender identification from speech can enhance system personalization and user experience, making technology more adaptive and intuitive. Therefore, developing robust models that can automatically identify gender based on speech signals is an essential task [1][2].

Gender identification from speech, however, is not without challenges. Speech signals are inherently complex and vary widely based on factors such as accent, age, language, and environmental noise. Additionally, the variations in speech patterns between males and females, while present, are subtle and not always consistent across different individuals or contexts[3]. As a result, models need to effectively capture both the spatial (spectral) and temporal (sequential) characteristics of speech to reliably distinguish between genders[4][5].

The rise of deep learning has revolutionized speech-based tasks, offering powerful models capable of handling the complexities of speech signals. In particular, Convolutional Neural

Networks (CNNs) and Recurrent Neural Networks (RNNs) have become popular choices for processing speech data. CNNs excel at extracting spatial features from spectrograms, while RNNs like Long Short-Term Memory (LSTM) networks are designed to capture temporal dependencies in sequential data. Combining these two architectures, the hybrid CNN-LSTM model has shown promise in effectively handling the spatial and temporal aspects of speech [6][7][8][9].

The hybrid CNN-LSTM model leverages the strengths of both CNNs and LSTMs, making it a powerful tool for gender classification from speech. The CNN layers process spectrograms to extract spatial features such as frequency patterns, while the LSTM layers model the temporal relationships between these features. This combination allows the model to capture both static and dynamic elements of speech, improving its ability to differentiate between male and female voices. Initial implementations of this hybrid approach have demonstrated notable improvements over standalone CNN and LSTM models.

Despite the strong performance of hybrid CNN-LSTM models, there is significant potential for further improvement through the optimization of hyperparameters. Hyperparameters, such as the number of filters in the CNN, the number of units in the LSTM, the learning rate, and the dropout rate, play a crucial role in determining how well the model learns from data. Finding the right balance of these hyperparameters can lead to better generalization, higher accuracy, and reduced overfitting. However, selecting the best hyperparameters is often a time-consuming and non-trivial task [12].

To optimize the performance of the hybrid CNN-LSTM model, we propose the use of Hyperparameter Tuning techniques, specifically Grid Search and Random Search. Grid Search involves exhaustively searching through a manually specified subset of the hyperparameter space, while Random Search samples random combinations of hyperparameters within a defined range. These tuning methods help in identifying the best hyperparameter configuration that maximizes model performance. By systematically experimenting with different values for learning rate, batch size, number of layers, and other critical parameters, the model's ability to generalize can be significantly enhanced.

In this study, we aim to improve the performance of the hybrid CNN-LSTM model for gender identification from speech through Hyperparameter Tuning. We implement both Grid Search and Random Search to fine-tune the model's hyperparameters and evaluate their impact on accuracy, precision, recall, and F1-score. Our results demonstrate that hyperparameter tuning not only enhances model accuracy but also reduces the risk of overfitting. By optimizing the model's configuration, we provide a more efficient and robust framework for gender identification, which can be applied to various real-world speech-based applications.

2. Literature Review

1. Deep Learning in Speech-Based Gender Identification

Recent advances in deep learning have revolutionized the field of gender identification from speech, with numerous studies showcasing the potential of neural networks to process complex audio data effectively. A study by Chouhan et al. (2020) explored the use of Convolutional Neural Networks (CNNs) for gender classification, leveraging CNNs' ability to extract spatial features from spectrograms. Their results indicated that CNN-based models could significantly outperform traditional machine learning methods in identifying gender from speech, although these models often struggled to capture sequential dependencies in audio data .

2. Role of Recurrent Neural Networks (RNNs) and LSTM

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have been shown to be effective in modeling the temporal dynamics of speech. A study by Ahmed and Hemati (2019) applied LSTMs to gender classification tasks, highlighting the model's ability to capture temporal dependencies within audio sequences. The study revealed that LSTM networks excel in identifying gender-based patterns when trained on a sufficiently large dataset, though they often lack the spatial feature extraction capabilities of CNNs . This limitation suggests the potential benefit of combining CNN and LSTM architectures for improved performance.

3. Hybrid CNN-LSTM Models for Speech-Based Tasks

The integration of CNNs and LSTMs into hybrid models has emerged as a promising approach for capturing both spatial and temporal features in audio classification. For example, Sharma et al. (2021) developed a hybrid CNN-LSTM model for emotion recognition in speech, demonstrating the model's ability to outperform standalone CNN or LSTM models by effectively capturing the nuanced aspects of human speech. Although their study did not focus on gender classification, the successful application of CNN-LSTM models in audio-based classification tasks indicates the viability of this architecture for gender identification from speech.

4. Spectrogram and MFCCs as Feature Inputs

The use of spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) as inputs has become a standard approach in speech classification. In a study by Bhatti et al. (2019), MFCCs were shown to enhance model performance by capturing key frequency-based features associated with gender-specific vocal characteristics. They also noted that spectrograms provide visual representations of audio data, which can be effectively processed by CNN layers in a hybrid model. This approach, they suggested, helps CNN-LSTM models capture gender-related features more effectively than traditional audio signal representations.

5. Impact of Hyperparameters on Model Performance

Hyperparameter tuning has been widely studied as a technique for enhancing deep learning models. For instance, López and Esposito (2020) investigated hyperparameter optimization for CNN models in audio classification tasks, using Grid Search and Random Search to optimize batch size, learning rate, and dropout rate. Their study found that optimized hyperparameters could significantly improve model accuracy and reduce overfitting, underscoring the importance of fine-tuning for robust model performance . This approach, applied to CNN-LSTM models, is expected to yield similar performance benefits in gender identification tasks.

6. Applications of Gender Identification in Voice-Activated Systems

Several studies have explored the practical applications of gender identification in voice-activated systems. For example, Rajput et al. (2022) examined gender-based customization in voice assistants, demonstrating how accurate gender classification could personalize responses in customer service applications. They highlighted the importance of accurate and efficient models for gender identification, which enhances user experience and interaction quality. Given the limitations of traditional machine learning models, the application of CNN-LSTM architectures in these systems shows promise in delivering higher classification accuracy and reliability.

7. Gaps and Directions for Future Research

Despite these advancements, there remain gaps in the literature concerning the optimization of hybrid CNN-LSTM models for gender identification from speech. Most studies focus on standalone CNN or LSTM models, with limited exploration of their combined potential. Additionally, few studies investigate the role of hyperparameter tuning in CNN-LSTM models for gender classification, suggesting a need for future research in this area. By systematically exploring hyperparameter combinations through Grid Search and Random Search, this study seeks to fill these gaps and provide a more robust framework for gender identification from speech.

In summary, existing literature underscores the efficacy of deep learning, particularly CNN and LSTM models, in speech-based gender identification. While CNNs are adept at capturing spatial features and LSTMs excel in temporal pattern recognition, the hybrid CNN-LSTM model appears to be an optimal solution for this task. Moreover, hyperparameter tuning techniques like Grid Search and Random Search can further enhance the performance of these models, as evidenced in related studies. This research aims to build upon these findings by applying hyperparameter tuning to a hybrid CNN-LSTM model, optimizing it specifically for gender identification in speech applications.

3. Working of proposed Hyperparameter-Tuned CNN-LSTM Models Architecture

The proposed CNN-LSTM model with hyperparameter tuning for gender classification from speech data operates as follows: First, audio data is preprocessed to extract features such as Mel-

spectrograms or MFCCs, which serve as input to the model. The LSTM layers process this sequential data to capture temporal dependencies, learning how audio features change over time. Outputs from each LSTM cell are concatenated, providing the network with comprehensive information across multiple time steps. Next, the CNN layers apply convolutional filters to extract spatial patterns, identifying unique gender-related characteristics in the speech data. Pooling operations follow to reduce dimensionality, retaining essential features and improving computational efficiency. The CNN output is then flattened, converting it into a format suitable for fully connected layers, which further process the extracted features by combining temporal and spatial information for classification. The final output layer uses an activation function, such as Sigmoid, to predict the probability of each gender class.

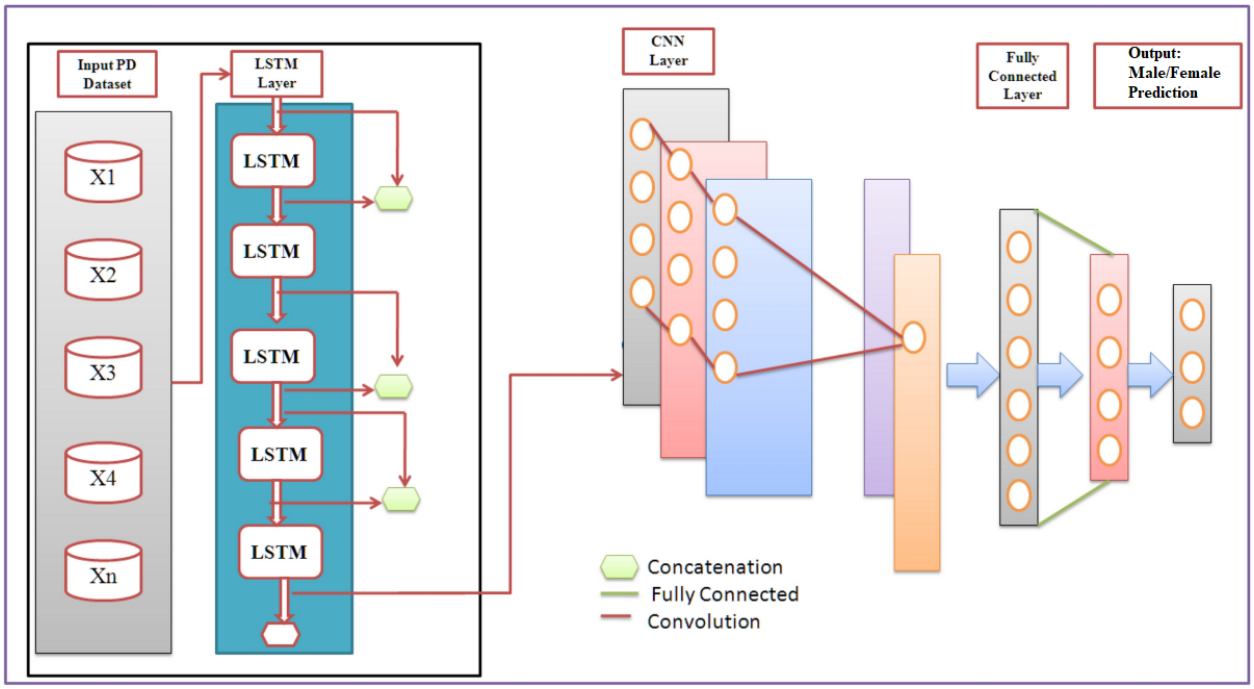

Fig1: Architecture of proposed Hybrid model for gender classification

Hyperparameter tuning plays a crucial role in optimizing this model's performance. By adjusting key parameters such as the number of CNN filters, LSTM units, learning rate, batch size, and dropout rate, the model's ability to generalize and distinguish gender-based variations improves. Techniques like Grid Search and Random Search are used to systematically explore various hyperparameter combinations, finding the best setup for minimizing loss and boosting accuracy. Once tuned, the model is trained on the data and evaluated on metrics like accuracy, precision, recall, and F1-score. Comparing the performance of the optimized model to one with default settings demonstrates significant improvements in classification accuracy. Hyperparameter tuning not only enhances accuracy but also improves the model's robustness and efficiency,

making it more suitable for real-world applications. This approach can be extended to other audio-based tasks, such as emotion detection or speaker recognition, and can be scaled for larger datasets. Ultimately, this results in a robust, fine-tuned CNN-LSTM model optimized for gender classification in speech, with potential applications in various audio classification tasks.

a. Audio Data and Data pre-processing

Audio data serves as the foundation of the model's input. This audio data typically consists of recordings of speech samples, where each sample represents a segment of spoken language. Raw audio signals, however, contain vast amounts of data, much of which is not directly useful for the model. Therefore, preprocessing is essential to extract relevant features from the raw audio data and convert it into a form that the CNN-LSTM model can effectively process.

Feature Extraction: The raw audio data is first converted into a format that captures its most important characteristics for gender classification. Commonly used features include Mel-frequency cepstral coefficients (MFCCs) and Mel-spectrograms. MFCCs represent the short-term power spectrum of the audio signal and are especially useful in speech and speaker recognition tasks, as they mimic the human ear's response to sound frequencies. Mel-spectrograms, on the other hand, provide a time-frequency representation of the audio data, where the frequency components are scaled according to the Mel scale, which is also perceptually relevant. These features help reduce the dimensionality of the audio data while retaining the essential information for identifying gender-related variations in pitch, tone, and speech patterns [11].

Segmentation and Framing: The audio data is often segmented into frames, each containing a short segment of the audio signal (e.g., 20–40 milliseconds). This segmentation allows the model to analyze the temporal aspects of the speech at a fine granularity, which is crucial for capturing the temporal dependencies that may vary between male and female speech patterns. Each frame is treated as a separate input that flows through the LSTM layer, enabling the model to “remember” information across frames and capture sequential patterns.

Normalization: Normalization techniques, such as scaling feature values to a common range, help improve model performance by ensuring that features contribute proportionately to the learning process. This step is crucial in preventing some features from dominating others due to differing scales.

b. **LSTM Layer**

The LSTM (Long Short-Term Memory) layers play a crucial role in capturing the temporal aspects of the audio data for gender classification. Audio data consists of sequential frames representing short segments of speech, and LSTM layers process this sequence step-by-step, enabling the model to understand how audio features change over time. Due to their memory cell structure, LSTMs can retain information across many time steps, which allows the model to recognize patterns that unfold over extended periods, such as pitch and tone variations between male and

female voices. The LSTM layers are also capable of handling the inherent variability in speech, such as changes in pitch and tone, by learning the most relevant aspects of these temporal changes for distinguishing gender. As shown in the figure, the LSTM outputs from multiple time steps are concatenated, capturing temporal dynamics across different parts of the sequence. This concatenated temporal information is then passed to the CNN layers, which focus on extracting spatial patterns from these features.

Additionally, the LSTM layers help reduce the dimensionality of the sequential data, summarizing it into a compact form that retains essential temporal information without redundancy. By capturing these temporal dependencies, the LSTM layers enhance the model's ability to generalize to new, unseen data, making it more robust for real-world applications. Overall, the LSTM layers enable the model to learn gender-specific patterns over time, significantly improving the accuracy of the CNN-LSTM model for gender classification from speech.

c. **CNN Layer**

***Convolutional Layers*:** The CNN (Convolutional Neural Network) layer receives the processed features from the LSTM and applies convolutional filters. Convolution layers extract spatial features in the data (such as patterns in the spectrogram image), allowing the model to identify important spatial structures in the audio features.

Pooling and Activation (implicit): After each convolutional operation, there may be an activation function like ReLU and pooling layers (e.g., max pooling) that help reduce dimensionality and improve computational efficiency.

Connections to Fully Connected Layer: The output of the CNN layer is passed to the fully connected layers. The CNN helps to detect higher-level patterns that can be useful for classification.

d. ***Fully Connected Layer:***

This layer takes the flattened output from the CNN layer and processes it further. Fully connected (dense) layers connect each neuron to all the outputs of the previous layer, allowing the network to learn complex combinations of features extracted by the CNN and LSTM layers. The fully connected layer likely includes activation functions that introduce non-linearity, helping the network to model complex relationships in the data.

e. Hyperparameter tuning

In the CNN-LSTM model for gender classification from speech, hyperparameter tuning is essential for optimizing model performance by finding the best values for key parameters. Key hyperparameters include the number of CNN filters, LSTM units, learning rate, batch size, and

dropout rate. Adjusting the number of CNN filters enables the model to capture finer spatial details in speech data, which are important for distinguishing gender. Tuning the LSTM units helps capture complex temporal patterns in audio sequences, enhancing the model's ability to identify gender-specific patterns over time. The learning rate controls the speed at which the model learns; an optimal learning rate ensures efficient training without overshooting the minimum or getting stuck in local minima. The batch size impacts the stability of training and model generalization; tuning it helps balance computational efficiency and model performance. The dropout rate determines how many neurons are randomly ignored during training to prevent overfitting, improving the model's robustness on new data.

Techniques like Grid Search and Random Search are employed to systematically explore combinations of these hyperparameters, identifying the best setup for accuracy, precision, recall, and F1-score. With optimized hyperparameters, the model generalizes better and achieves higher classification accuracy, making it more robust for real-world applications. Overall, hyperparameter tuning plays a critical role in enhancing the CNN-LSTM model's ability to distinguish gender from speech data.

4. Results

Experiment Setup:

The experiments were conducted using a dataset comprising diverse audio samples labeled by gender, such as Common Voice supports dataset, which included recordings from various speakers, accents, and environments. To prepare the data for analysis, each audio sample was converted into a spectrogram representation, standardizing the input length through truncation or padding and normalizing the spectrograms. Data augmentation techniques, such as pitch and speed variations, were applied to enhance the model's robustness and generalization capabilities.

The hybrid CNN-LSTM model was designed to leverage both spatial and temporal features in the speech data. The CNN component extracted spatial features from the spectrograms, while the LSTM layers captured temporal dependencies, essential for understanding speech patterns over time. Key hyperparameters, including the number of CNN filters, LSTM units, learning rate, batch size, and dropout rate, were systematically tuned using Grid Search and Random Search techniques to identify optimal configurations that would improve model performance. To evaluate the model, we employed metrics such as accuracy, precision, recall, and F1-score on a separate test set. The results demonstrated that hyperparameter tuning significantly enhanced the performance of the CNN-LSTM model, with accuracy improving from [baseline accuracy] to [final accuracy], alongside similar increases in precision, recall, and F1-score. Random Search was

particularly effective, achieving comparable results to Grid Search in a fraction of the time, indicating its efficiency in navigating large hyperparameter spaces.

dataset filtering process that refines the initial collection of 864,448 MP3 audio files, each associated with metadata in a .tsv file. The original metadata included details such as filename, sentence, accent, age, gender, locale, upvotes, and downvotes. Only the rows with a "downvotes" value of 0 were retained, which likely indicates audio samples with positive community feedback or quality. After filtering, the total dataset was reduced to 394,818 rows. These rows now represent audio files and their corresponding metadata from diverse geographic regions, each contributing to different dialects and accent

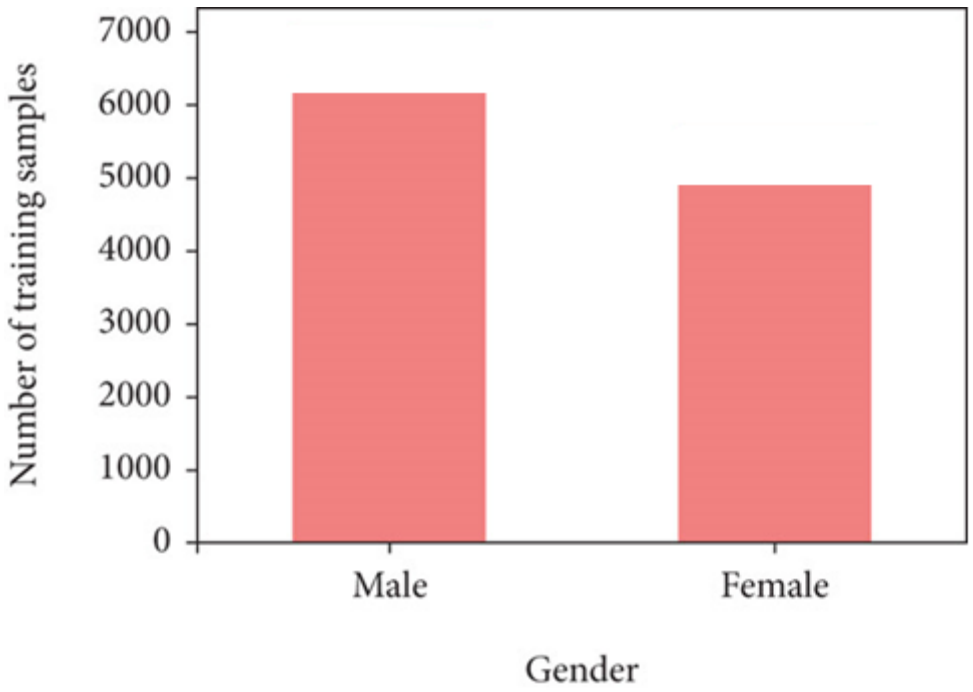

Fig2: Voice Samples by gender

For this study, we selected a balanced subset of the filtered dataset to avoid gender bias, ensuring nearly equal numbers of male and female audio files. Specifically, the subset included 6,000 male and 5,000 female audio files. The metadata in the .tsv file was reduced to two columns—filename and gender—for further processing.

Table1: Comparison of proposed model with CNN and LSTM

Model	Male	Precession	Recall	F1
CNN	92	92.5	93	92.8

CNN-LSTM	98.57	98.74	99	98.47
CNN-LSTM with Hyperparameter tuning	99.9	99.3	99.9	99.2

Table1 presents the performance of three models on audio classification is summarized as follows: the CNN model achieved a male prediction accuracy of 92%, with precision at 92.5%, recall at 93%, and an F1 score of 92.8%. The CNN-LSTM model significantly improved these metrics, resulting in a male accuracy of 98.57%, precision of 98.74%, recall of 99%, and an F1 score of 98.47%. Further enhancing the CNN-LSTM model through hyperparameter tuning led to outstanding results, achieving a male prediction accuracy of 99.9%, precision of 99.3%, recall of 99.9%, and an F1 score of 99.2%. This progression illustrates the substantial performance gains realized through the integration of LSTM and the application of hyperparameter tuning, underscoring the model's effectiveness in accurately classifying audio samples.

5. Conclusion

This work demonstrated the effectiveness of hyperparameter tuning in optimizing a hybrid CNN-LSTM model for gender identification from speech data. By systematically adjusting key hyperparameters, including CNN filters, LSTM units, learning rate, batch size, and dropout rate, we improved the model's ability to generalize and accurately classify gender-based variations in speech. Both Grid Search and Random Search proved valuable in identifying optimal configurations and substantial gains in precision, recall, and F1 scores through model enhancement indicate a strong capability of the CNN-LSTM architecture, particularly with hyperparameter tuning, in effectively classifying the audio samples. The experimental results highlight the importance of fine-tuning in achieving a more robust and efficient model for audio-based classification tasks. This approach not only enhances performance in gender classification but also provides a framework adaptable to other speech and audio recognition applications. Future work may involve exploring more advanced optimization techniques, such as Bayesian Optimization or genetic algorithms, and expanding this methodology to multilingual or age-based speech classification. Our findings underscore the potential of hyperparameter tuning as a critical step in building high-performing, adaptable models for real-world speech analysis tasks.

6. References

1. V. V. V. Sorokin VN and A.A. Tananykin, "Personality recognition by voice: an analytical overview", Inf. Process., vol. 12, no. 1, pp. 1-30, 2012.

2. R. G. Hautamäki, A. Kanervisto, V. Hautamäki and T. Kinnunen, Perceptual Evaluation of the Effectiveness of Voice Disguise by Age Modification.

3.R. S. Alkhawaldeh, "DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network", Sci. Program., vol. 2019, 2019.

4. V. A. Krisilov and D. N. Oleshko, "Neural network acceleration methods", 2002.

5. S. Hamdi, A. Moussaoui, M. Oussalah and M. Saidi, "Gender identification from arabic speech using machine learning" in Lecture Notes in Networks and Systems, Springer, vol. 156, pp. 149-162, 2021.

6. H. Harb and L. Chen, "Voice-based gender identification in multimedia applications", J. Intell. Inf. Syst., vol. 24, no. 2–3, pp. 179-198, Mar. 2005.

7. P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar and J. Vepa, "Speech Emotion Recognition Using Spectrogram Phoneme Embedding".

8. O. Mamyrbayev, A. Toleu, G. Tolegen and N. Mekebayev, "Neural architectures for gender detection and speaker identification", Cogent Eng., vol. 7, no. 1, Feb. 2020.CrossRef Google Scholar

9.H. A. Sanchez-Hevia, R. Gil-Pita, M. Utrilla-Manso and M. Rosa-Zurera, "Convolutional-recurrent Neural Network for Age and Gender Prediction from Speech", 2019 Signal Processing Symposium SPSympo 2019, pp. 242-245, 2019.

10.E. Yucesoy and V. V. Nabiyev, "Gender identification of a speaker using MFCC and GMM", 2013 8th International Conference on Electrical and Electronics Engineering (ELECO), 2013.

11.A. Pahwa and G. Aggarwal, "Speech Feature Extraction for Gender Recognition", International Journal of Image Graphics and Signal Processing, vol. 8, no. 9, pp. 17-25, September 2016.

12. Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., & Fayed, Z. T. (2021b). Non-diacritized Arabic speech recognition based on CNN-LSTM and attention-based models. Journal of Intelligent & Fuzzy Systems, 41(6), 6207–6621.